Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel

Bernard Charlin,[1] Robert Gagnon,[1] Jean Pelletier,[2] Michel Coletti,[3] Grace Abi-Rizk,[4] Claudine Nasr,[4] Évelyne Sauvé[1] & Cees van der Vleuten[5]

CONTEXT The Script Concordance Test (SCT) assesses reasoning in the context of uncertainty. Because there is no single correct answer, scoring is based on the comparison of answers provided by examinees with those provided by members of a reference panel made up of experienced practitioners. The study aimed to assess the discriminatory power of the SCT based on the variability of the reference panel's answers.

METHOD Items from a bank covering different family medicine domains were classified into 3 groups according to the degree of variability of answers provided by a pool of experienced doctors. A variability index (mean squared error) was used to select items in the low, moderate and high variability categories. A 102-item test (Cronbach's α 0.70), made up of 3 subtests of each category, was administered to 3 contrasting groups in family medicine: 157 clerkship students, 30 residents and 30 practising doctors. ANOVA and effect size (ES) were used to quantify and test the discrimination power of the 3 subtests.

RESULTS The high variability subtest showed high effect size for discrimination between extreme groups (ES = 1.5; $F = 16.3$, $P < 0.001$), whereas the moderate variability subtest showed less effect size (ES = 0.56; $F = 57$, $P = 0.041$). The low variability subtest did not discriminate significantly (ES = 0.31; $F = 2.9$, $P = 0.06$).

CONCLUSIONS Variability of answers within the reference panel is a key component of the discriminatory power of the SCT. In accordance with theory, the presence of variability ensures discrimination between levels of clinical experience. These results imply important considerations for the construction of efficient SCTs.

KEYWORDS family practice/*education; clinical competence/*standards; *uncertainty; reference values; *decision making; psychological tests/*standards; analysis of variance.

Medical Education 2006; ***40****: 848–854*
doi:10.1111/j.1365-2929.2006.02541.x

INTRODUCTION

A significant part of professionals' competence relies on the capacity to deal with uncertainty[1] and to solve ill-defined problems.[2] Traditional written assessment tools, such as rich-context, multiple-choice questions, properly and reliably test the ability of students to apply well known solutions to well defined problems. However, assessment of reasoning competence should also include tools that measure the ability to rationally solve ill-defined problems. The Script Concordance Test (SCT)[3] was developed with the purpose of expanding the material assessed in clinical reasoning to include ill-defined problems. It is designed to be added to existing tools, not substituted for them.

The name of the test, script concordance, reflects the theory on which it is based.[4,5] The test consists of a

[1] URDESS - Faculty of Medicine, University of Montreal, Montreal, Quebec, Canada
[2] Department of Family Medicine, Montreal, Quebec, Canada
[3] Department of Family Medicine, University of Bobigny, Bobigny, France
[4] Department of Family Medicine, University Saint-Joseph, Beirut, Lebanon
[5] Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands

Correspondence: Bernard Charlin MD, PhD, URDESS - Faculty of Medicine, University of Montreal, CP 6128, Succursale Centre-Ville, Montreal, Quebec H3C 3J7, Canada. Tel: 00 1 514 343 7827; Fax: 00 1 514 343 7650; E-mail: bernard.charlin@umontreal.ca

Overview

What is already known on this subject

The Script Concordance Test (SCT) reliably assesses reasoning in the context of uncertainty. Scoring is based on a comparison of answers provided by examinees with those provided by members of a reference panel made up of experienced practitioners. The relationship between the variability of the reference panel's answers and the discriminative power of SCTs according to levels of experience is not known.

What this study adds

In accordance with theory, variability within answers provided by the reference panel is a key component of the discriminative power of the SCT.

Suggestions for further research

Test designers wishing to discriminate examinees by clinical experience (and clinical reasoning competence) should deliberately build items that will induce variability within reference panel answers.

series of challenging authentic clinical situations described in vignettes (paper or computer-based). Items are derived from the questions and actions doctors actually ask and make when they are confronted with such situations in clinical practice. Several options (diagnosis, management or attitude) are relevant. In accordance with what is known about clinical reasoning processes,[6–9] a Likert scale, measuring the qualitative judgements that are iteratively made during these processes, captures examinees' answers (Table 1). The method for building tools according to these principles has been described in detail elsewhere.[10]

Assessment on ill-defined problems is difficult because professionals in similar situations do not collect exactly the same data and do not follow the same paths of thought.[9] Professionals also show substantial variation in performance on any particular real or simulated case.[7,9,11] Any assessment system has to take this variability into account. The SCT uses an aggregate scoring method that reflects the variability experts demonstrate when they reason in clinical situations.[12,13] Scores on each item are derived from the answers given by members of a reference panel made up of doctors experienced in the domain.

Usual models of analysis cannot be applied to the SCT, which by design includes several credit points for different answers to an item. In a previous paper[11] we showed that if panel members are asked to provide the 'right answers' to be expected from students, they tend to change the answers they spontaneously provide when they are placed in problem-solving situations in the same contexts as examinees. The inadequacies of traditional models for assessment in the context of uncertainty bring challenges, one of which concerns how much variability is advisable in a test. In traditional models, an item with discrepancies in the answers given by a reference panel is considered to be of bad quality, while in the SCT approach a reasonable amount of variability may contribute to the discriminative power of the item.

The implications of using this variability as a way of detecting levels of clinical experience in a population of examinees is therefore an important research issue: what is the 'signal' that adds to the incremental validity of the measurement? Our goal was to study formally the amount of variability that optimises the discriminative power of items and the discriminative power of the test as a whole. Our research question was: Does variation in the answers of members of reference panels used to establish an aggregate scoring key influence the ability of the test to differentiate between levels of clinical experience?

From empirical research[11,14,15] and from Schön's model of professional practice,[2] we assumed that clinical experience is related to the capacity to reason on ill-defined problems and hence that items with variability would have a discriminative power to detect experience. Items with consensus (low variability) among panel members would probe well defined problems and would have less discriminative power. By contrast, we also know that even if experienced clinicians diverge on detail in their way toward solutions to clinical problems, they generally agree on reasoning outcomes. Therefore, items with large margins of disagreement would not be good indicators of clinical experience. These items reflect measurement error (noise) and are not useful for assessment. Expressed in terms of effect size, we hypothesised a relationship (inverse U-shape

Table 1 Example of a clinical vignette and format of items used for diagnostic knowledge assessment

A 17-year-old girl arrives suffering from dyspnea. She is out of breath and has been brought very rapidly to your office. At the moment of onset of dyspnea, she was in a car returning home after a desensitization injection at her doctor's clinic

If you were thinking of:	And the patient reports or you find upon clinical examination	This hypothesis becomes:				
Anaphylactic reaction	Respiratory rhythm at 32	− 2	− 1	0	+ 1	+ 2
Asthma	Difficulty swallowing	− 2	− 1	0	+ 1	+ 2
Hyperventilation	A normal pharynx	− 2	− 1	0	+ 1	+ 2
Anaphylactic reaction	Arterial blood pressure = 120/180	− 2	− 1	0	+ 1	+ 2
Asthma	A diffuse arterial II/IV murmur	− 2	− 1	0	+ 1	+ 2
Hyperventilation	Arterial blood pressure = 150/90	− 2	− 1	0	+ 1	+ 2

− 2 = ruled out or almost ruled out; − 1 = less probable; 0 = neither less nor more probable; + 1 = more probable; + 2 = certain or almost certain.

relation) between effect size and level of clinical experience, with moderate variability associated with maximal effect size and extreme variability (low and high) associated with minimal effect size.

METHODS

Test material

A bank of 145 items pertaining to family medicine was developed by a team composed of experienced family doctors in charge of maintaining standards of practice at the Quebec Board of Physicians.

In order to identify items to which experienced doctors would give answers with different levels of variability, a group of 51 family doctors (13 from Quebec and 38 from France) were asked to answer the 145 items. The variability of their responses was addressed by using the mean square error (MSE). When items are measured on the same scale, the MSE allows a variability estimate that is comparable between items to be obtained.[16] In order to obtain interpretable coefficients, the original scoring scheme of − 2, − 1, 0, + 1, + 2 was transformed to scores of 1–5.

The final test used in the study was built with 2 requirements. It had to be answerable in a realistic amount of time (about 1 hour) and it had to include balanced subsets of items defined by their levels of variability. A third of the items were chosen in the low variability range (MSE 0.0–0.49), a third in the middle variability range (MSE 0.50–0.99), and a third in the high variability range (MSE ≥ 1.0). The final test therefore had 102 items, with 34 items in each of the 3 ranges. Item selection stopped when there was 34 items in each group. The remaining 43 items did not differ from those used in the final test.

Respondents

The detection of levels of clinical experience was studied on 3 levels: clerkship students (157), residents (Years 1 and 2 for a total of 30) and faculty staff (30). The latter were new respondents and had not participated in the item selection phase. Respondents were recruited from the Department of Family Medicine at the Faculty of Medicine, University of Montreal. All participated freely in the study and signed informed consent.

Scoring process

The SCT technique requires the recruitment of a panel of reference. This panel consists of experienced practitioners whose presence on a jury is legitimate with regard to the level of the persons assessed. Panel members are asked to fill out the test exactly as examinees will do and their answers are then used to constitute the scoring key.

The 30 staff members were used both as examinees (third level of clinical experience) and as a panel of reference. They took the test individually. The answers obtained served to construct the scoring key, according to the aggregate methodology used in the SCT.[10] For each item, each examinee answer received a credit mark corresponding to the proportion of panel members that selected it. The maximum score for each item was 1 for the modal answer. Other panel members' choices received a partial credit. Answers not chosen by panel members received 0. To get this proportional transformation, the number of members who provided an answer on the Likert scale was divided by the modal value for the item. If, for example, on a given item, 20 members (out of 30) chose response 1 on the Likert scale, this choice received the maximum score of 1 point (20/20). Then, if 8 members chose response 2, this choice received 0.4 (8/20). Finally, if 5 panel members chose response 0, this choice received 0.25 (5/20). The total score for the test was the sum of credits obtained on each item, which in the end was transformed to obtain a maximum of 100.

Because we did not want staff members having their own answers used in the scoring process (and thus being subjected to a positive scoring bias), their scores on each item were computed with a scoring key that excluded each person's own answer. Thus, for this group, scoring was carried out with scoring keys derived from the answers given by the other 29 members of the group. The answers of respondents at the other 2 levels were computed with the same scoring key (made of 30 panel members).

Statistical analysis

Performance for each subtest was calculated by summing the scores of its items and transforming that value into a percentage. When values were missing, they were replaced by a score of 0, assuming no concordance. Almost all respondents (95%) had 4 or fewer missing values. Data from 1 student with 65 missing values were excluded from the analysis.

To study the effect of variability on the discrimination of subtests along the level of clinical experience, a repeated-measure ANOVA was used with the 3-variability category as a within-subject factor. Effect size (ES) was used as an index to quantify the magnitude of the difference between the groups at subtest level and item level. According to Cohen,[17] a small effect size is 0.2, a medium effect size is 0.5 and a large effect size is 0.8. All tests were 2-tailed and a *P*-value < 5% was considered statistically significant.

RESULTS

Cronbach's α values were 0.70 for the whole test, 0.46 for the low variability subtest, 0.51 for the moderate variability subtest, and 0.40 for the high variability subtest.

Analyses of the discriminative power of the 3 subtests on the different levels of clinical experience (Fig. 1) indicate that the low variability subtest yielded the highest scores: the students' mean score was 74.4 (SD 7.7); the residents' mean score was 77.8 (SD 5.1), and the staff members' mean score was 76.4 (SD 9.6). Note that the staff members' mean score was lower on this subtest than that of the residents. The students' performance was close to those of residents and staff, and no discrimination between groups was observed with this subtest ($F = 2.78$, $P = 0.06$). The high variability subtest yielded the lowest mean scores for each group: 63.3 (6.8), 70.5 (7.6), 72.0 (5.2). This subtest discriminates along levels of experience ($F = 31.28$, $P < 0.001$). The moderate variability subtest yielded intermediate mean scores: 68.5 (7.5), 72.2 (8.9), 72.6 (7.2) and statistically significant discriminating power ($F = 5.67$, $P = 0.001$).

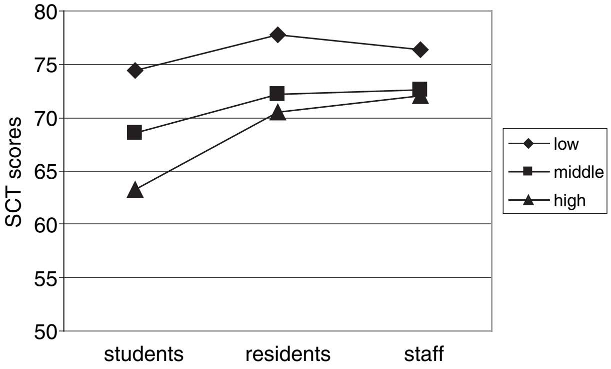

Figure 1 Performance of groups of different levels of clinical experience on the 3 subtests (mean of groups per category of variability).

Analyses at group level show a statistically significant difference in performance on the 3 subtests for students ($F = 146.0$, $P < 0.001$), residents ($F = 15.6$; $P < 0.001$) and staff members ($F = 5.3$, $P = 0.01$). The 2-factor repeated measure of variance shows a clear effect of variability ($F = 52.7$, $P < 0.001$) and a significant interaction between level of variability and level of clinical experience ($F = 5.1$, $P < 0.001$).

On analyses of effect size at the group level, the high variability subtest showed high effect size for discrimination between extreme groups (ES = 1.45; $F = 16.3$, $P < 0.001$), while the moderate variability subtest showed less effect size (ES = 0.56; $F = 57$, $P = 0.041$). The low variability subtest did not discriminate significantly (ES = 0.23; $F = 2.9$, $P = 0.06$). Figure 2 pictures the effect size for staff–students and staff–residents differences. The expected relationship (inverse U-shape relation) between effect size and level of clinical experience was not found. Instead, there was a clear tendency towards greater difference and better discrimination with greater levels of item variability. Effect size is far greater between staff and students than between staff and residents.

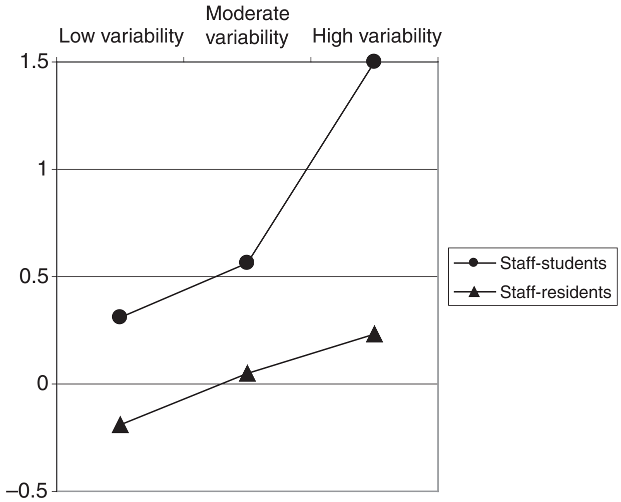

Figure 2 Subtest capacity to discriminate experienced doctors (staff members) from students and residents. The high variability subtest clearly has clearly the better discriminative power.

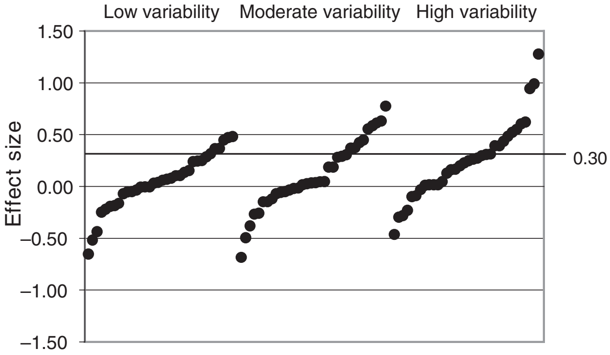

Figure 3 Individual effect sizes for each item in the 3 variability categories (students versus staff). The percentage of items with effect size > 0.30 decreases from high variability (38%) and moderate variability (26%) to low variability (18%).

Figure 3 shows the relationship between effect size (students versus faculty staff) and variability at the item level. For high variability items, 13 of 34 items have ES > 0.30 (38%). For moderate variability items, 9 of 34 items have ES > 0.30 (26%), and for low variability items only 6 of 34 (18%) have ES > 0.30.

DISCUSSION

The study is built on several assumptions.

1 Clinical experience is associated with better clinical reasoning, particularly better reasoning on ill-defined problems.
2 Reasoning on ill-defined problems is associated with variability in paths of thought among experienced clinicians.
3 A tool designed to assess reasoning on ill-defined problems should better detect clinical experience with items that generate variable answers among panel members than with items with low variability.
4 Items with high variability correspond to item difficulty related to uncertainty, whereas extreme variability reflects measurement errors rather than a signal.

The data did not show the expected relationship (inverse U-shape relation) between effect size and level of clinical experience. We found instead a monotonic relationship between item variability and level of clinical experience. Discrimination increased with variability. Effect size (i.e. large differences between groups) was substantial for both moderate and high variability subtests.

We explain this monotonic relationship by the absence of items with extreme variability (disagreement) (i.e. the total or near total dispersion of an answer along the 5 points of the scale). This implies a poor-quality item. Our item production phase did not produce such items and such items could not in any case be included in a test because they lack any face validity.

All groups obtained their higher mean score on the low variability subtest. Performances by the different groups were close for this subtest and no statistically significant discrimination of groups was observed. Residents on this subtest had a mean score higher than that of staff members. This 'intermediate effect' (students of intermediate level of experience get higher scores than experienced clinicians)[18] was not found with moderate and high variability subtests. According to the theory, this effect reflects the fact that low variability items assess formal clinical knowledge rather than judgement related to the integration of complex and ambiguous clinical data. We can speculate that staff members may possess more capacity than residents for judgement in complex situations, whereas formal clinical knowledge is less salient to them. The lowest mean scores were found with the high variability subtest, whereas results were intermediate with the moderate variability subtest.

The total test had a reasonable coefficient of reliability at 0.70 (Cronbach's α coefficient of internal coherence). It was noticeable that the moderate variability subtest had the highest reliability coefficient (0.51) in comparison with low (0.46) and high

(0.40) variability scales. Although the differences were not high, data indicate that, on a psychometric basis, moderate variability items show more interesting properties in terms of reliability (discrimination at the individual level), whereas high variability items show better discrimination at the group level of clinical experience.

Mean scores decreased with increasing variability. Between levels of low and high variability there was a loss of almost 11 points for students, a loss of 7 points for residents and a loss of 4 points for experienced doctors. Variability of items clearly had more impact on students than on residents or staff members. When there was variability in an item, experienced clinicians continued to show concordance with panel members, while students did not. This implies that variability of items should be taken into consideration when standards are being set for the SCT.

It must be acknowledged that our operationalisation of low variability may be biased by the fact that many items in this category (37%) used the extreme answer points of the scale (−2 or +2) and the variability of the answers was limited by the nature of the scale. This means that, if we had used a wider scale (such as a 7-point scale), some of these items may have had larger variability and may have been classified within the moderate variability group. This nevertheless does not minimise the results of the study. In fact, with a wider scale, the difference in metric quality of low variability versus higher variability may be larger than that we obtained.

The results of this study have important implications for the theoretical assumptions of the SCT. They confirm the tool's capacity to assess reasoning on ill-defined problems. They also confirm that if the goal of test designers is to discriminate examinees by clinical experience (and clinical reasoning competence) it is necessary to deliberately build items that will induce variability within the answers given by the reference panel. The variability of panel answers apparently contributes to the incremental validity of the SCT approach. Nevertheless, experience shows that it may be useful to place within SCTs items on which panel members agree (low variability items). These items assess the knowledge of well established solutions on well defined problems. This kind of item is close to the rich-context, multiple-choice question, but the item format and the task required of examinees both differ. We are currently in the process of comparing the measurement properties of these 2 types of test format.

Contributors: BC and RG contributed to the conception and design of the study, analysis and interpretation of data, and the writing of the paper. JP and MC contributed to the collection of data. GA-R and CN contributed to the conception of the study material. ES contributed to the collection and entry of data. CvdV contributed to the interpretation of data. All authors assisted in drafting the article and critical revision and approved the final version.
Acknowledgements: the authors thank François Goulet and André Jacques, Collège des Médecins du Quebec, who contributed to the conception of study material.
Funding: this research was funded by a grant from the National Board of Medical Examiners, Philadelphia, USA.
Conflicts of interest: none.
Ethical approval: the study was approved by the Ethical Board of the Faculty of Medicine, University of Montreal.

REFERENCES

1 Fox R. Medical uncertainty revisited. In: Albrecht G, Fitzpatrick R, Scrimshaw S, eds. *Handbook of Social Studies in Health and Medicine.* London: Sage Publications 2000;409–25.

2 Schön D. *The Reflective Practitioner: How Professionals Think in Action.* New York: Basic Books 1983.

3 Charlin B, van der Vleuten C. Standardised assessment of reasoning in contexts of uncertainty: the script concordance approach. *Eval Health Prof* 2004;**27**(3):304–19.

4 Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: theory and implication. *Acad Med* 1990;**65**(10):611–21.

5 Charlin B, Tardif J, Boshuizen HPA. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. *Acad Med* 2000;**75**:182–90.

6 Barrows HS, Norman GR, Neufeld VR, Feightner JW. The clinical reasoning of randomly selected physicians in general medical practice. *Clin Invest Med* 1982;**5**(1):49–55.

7 Elstein AS, Shulman LS, Sprafka SA. *Medical Problem Solving: An Analysis of Clinical Reasoning.* Cambridge, Massachusetts: Harvard University Press 1978.

8 Kassirer JP. Teaching clinical medicine by iterative hypothesis testing. Let's preach what we practise. *N Engl J Med* 1983;**309**(15):921–3.

9 Grant J, Marsden P. Primary knowledge, medical education and consultant expertise. *Med Educ* 1988;**22**:173–9.

10 Charlin B, Roy L, Brailovsky C, van der Vleuten C. The Script Concordance Test, a tool to assess the reflective clinician. *Teach Learn Med* 2000;**12**:189–95.

11 Charlin B, Desaulniers M, Gagnon R, Blouin D, van der Vleuten C. Comparison of an aggregate scoring

method with a consensus scoring method in a measure of clinical reasoning capacity. *Teach Learn Med* 2002;**14** (3):150–6.

12 Norman GR. Objective measurement of clinical performance. *Med Educ* 1985;**19** (1):43–7.

13 Norcini JJ, Shea JA. The use of the aggregate scoring for a recertification examination. *Eval Health Professions* 1990;**13**:241–51.

14 Brazeau-Lamontagne L, Charlin B, Gagnon R, Samson L, van der Vleuten C. Measurement of perception and interpretation skills along radiology training: utility of the script concordance approach. *Med Teacher* 2004;**26**:326–32.

15 Gagnon R, Charlin B, Coletti M, Sauvé E, van der Vleuten C. Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? *Med Educ* 2005;**39**:284–91.

16 Taylor SL, Payton ME, Raun WR. Relationship Between Mean Yield, Coefficient of Variation, Mean Square Error and Plot Size in Wheat Field Experiments. 1999. http://nue.okstate.edu/Index_Publications/MSE_Taylor_1999htm. [Accessed 24 July 2006.]

17 Cohen J. *Statistical Power Analysis for the Behavioural Sciences*. Hillsdale, New Jersey: Lawrence Earlbaum Associates 1988.

18 van der Vleuten CPM. The assessment of professional competence: development, research and practical implications. *Adv Health Sci Educ* 1996;**1**:41–67.

Received 14 April 2005; editorial comments to authors 23 August 2005; accepted for publication 1 November 2005